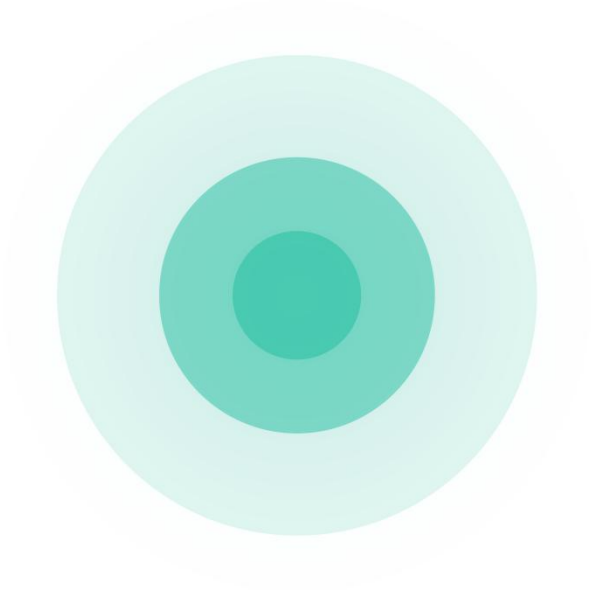


WARSAW DATA SCIENCE MEETUP

09.05.2017

Wykorzystanie zbiorów rozmytych w silnikach rekomendacji

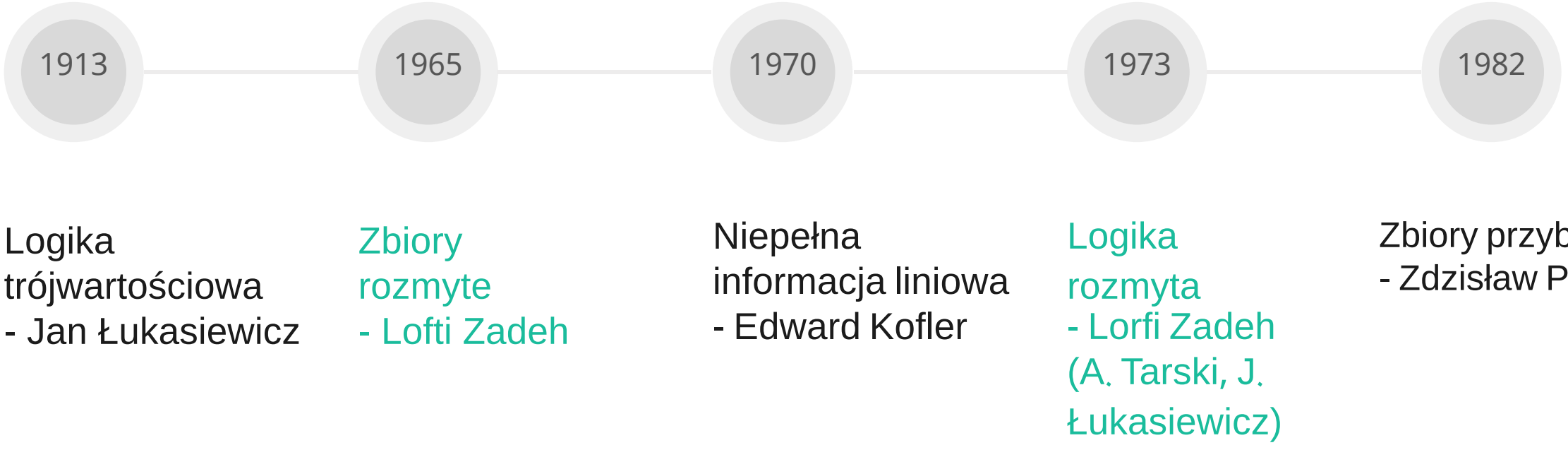
Mateusz Grzyb
konsultant technologiczny Microsoft Polska
mateuszgrzyb.pl

O czym będzie ta prezentacja?

Plan prezentacji

1. Zbiory rozmyte.
2. Logika rozmyta.
3. Systemy rekomendacyjne.
4. Przykład silnika rekomendacji wykorzystującego zbiory rozmyte.
5. Pytania do was.
6. Pytania do mnie.

"Miękkie" modelowanie - historia



Zbiory rozmyte

Zbiory rozmyte

Lofti Zadeh - 1965.
Rozszerzenie klasycznego zbioru z teorii zbiorów.
Obiekt matematyczny o zdefiniowanej funkcji przynależności.
Każdy element zbioru przyjmuje wartości z przedziału [0,1].
Każdy element zbioru, to dwójka uporządkowana.
Zastępuje logikę dwuwartościową logiką wielowartościową.

Umożliwia wykonywanie klasycznych operacji na zbiorach (suma, iloczyn, dopełnienie, etc.).

Zbiory rozmyte - zapis matematyczny

$$Z = \{(x, \mu(x))\}$$
$$gdzie : \mu : X \rightarrow [1,0]$$

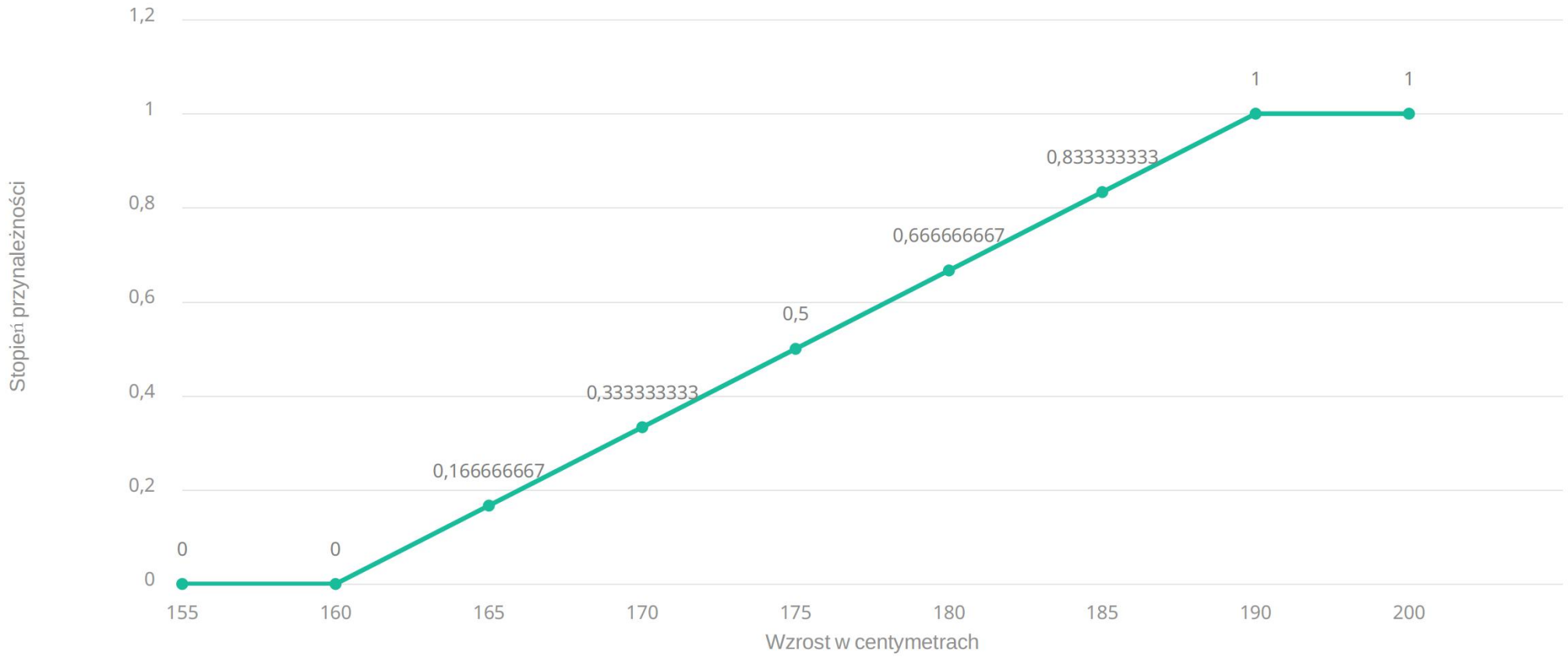
Czy mężczyzna o wzroście 185 cm jest wysoki, czy niski?

185 cm - wysoki czy niski?

$$\mu(x) = \frac{(wzrost - n)}{(w - n)}$$

gdzie :
n - osoba bezwzględnie niska
w - osoba bezwzględnie wysoka

Losowy zbiór mężczyzn

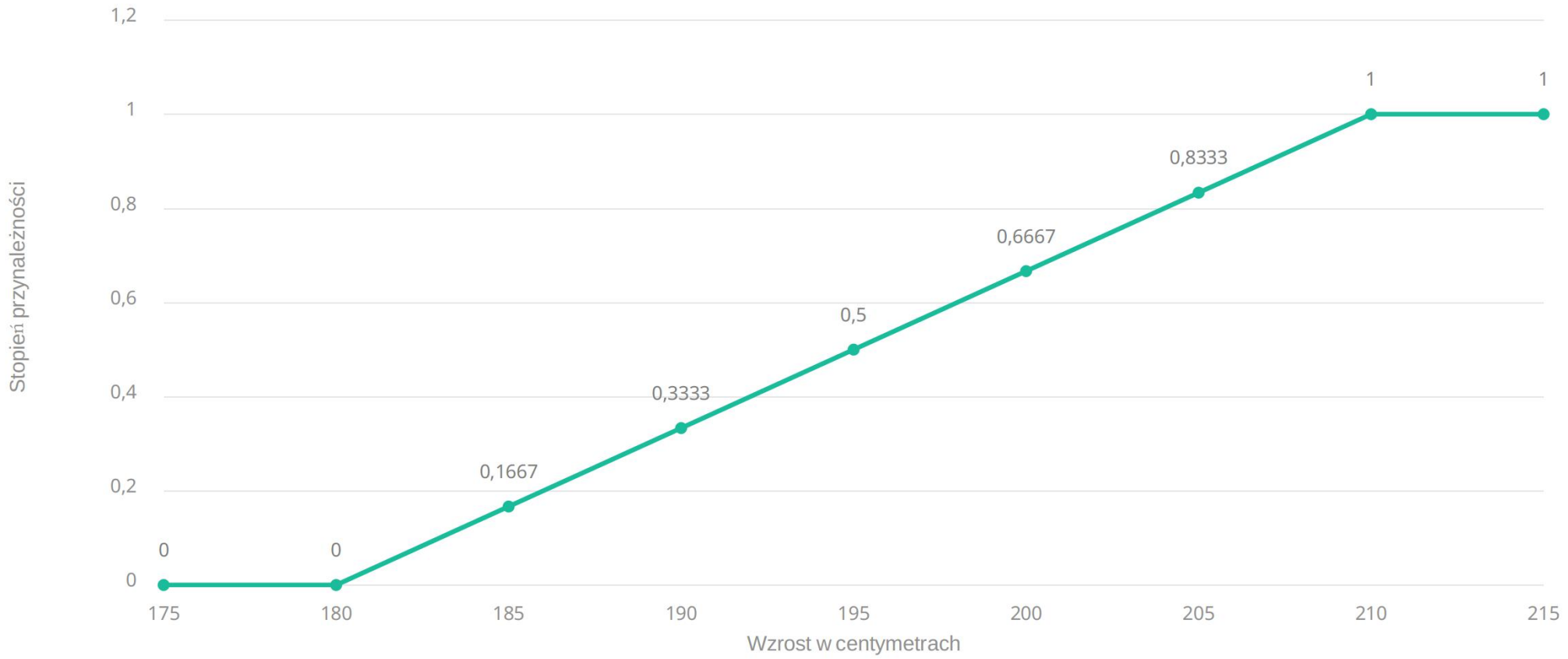


Losowy zbiór mężczyzn

$$\mu(x) = \frac{(185-160)}{(190-160)} = \frac{25}{30} = 0.833$$

Z = {(160, 0), (165, 0.167), (170, 0.333),
(175, 0.5), (180, 0.667), (185, 0.833), (190, 1)}

Losowy zbiór koszykarzy ligi NBA



Losowy zbiór koszykarzy ligi NBA

$$\mu(x) = \frac{(185-180)}{(210-180)} = \frac{5}{30} = 0.167$$

Z = {(180, 0), (185, 0.167), (190, 0.333),
(195, 0.5), (200, 0.667), (205, 0.833), (210, 1)}

Co warto zapamiętać z teorii zbiorów rozmytych?

Co warto zapamiętać z teorii zbiorów rozmytych?

- Istnieje wiele pośrednich stopni prawdy.
- Nie istnieją tu pojęcia prawdopodobieństwa i szansy.
- Każdy element zbioru to tzw. dwójka uporządkowana.
- Każdy element może przynależeć do zbioru z dowolną wartością stopnia przynależności z przedziału [0,1].

Logika rozmyta

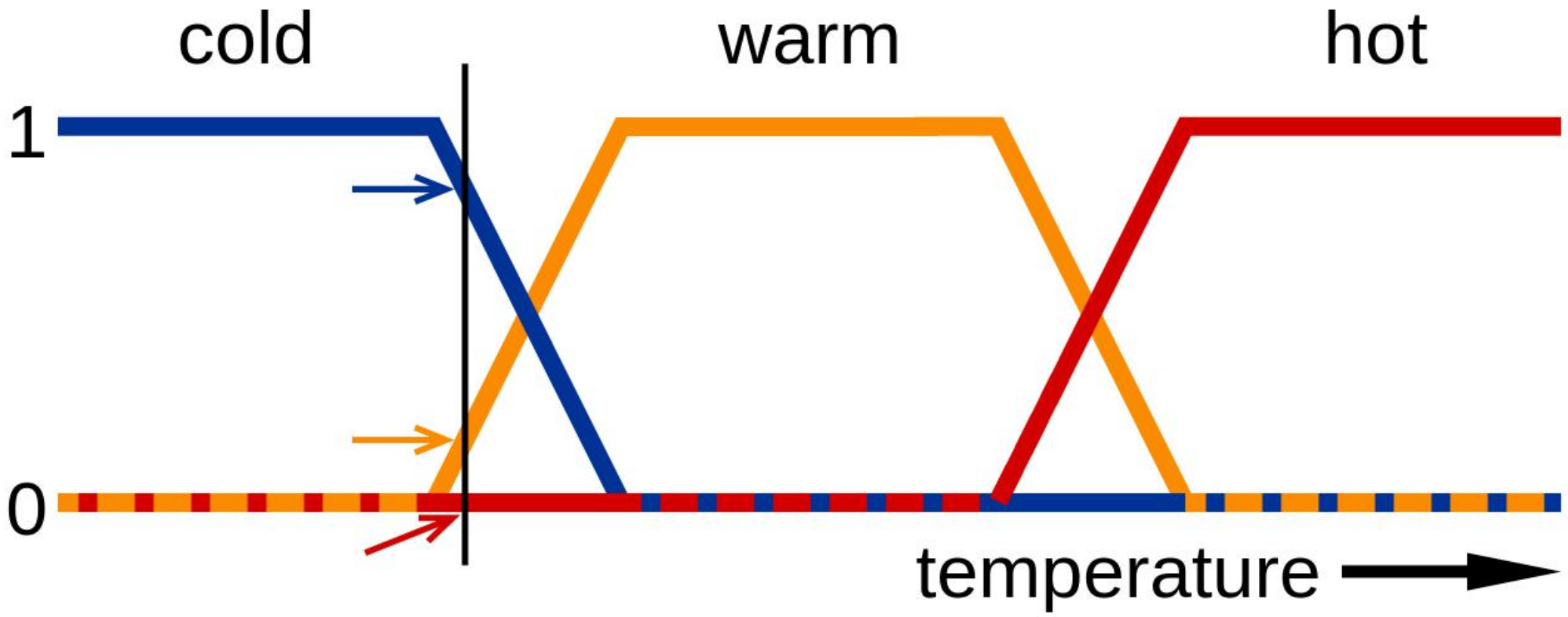
Logika rozmyta

Lofti Zadeh - 1973.

Ściśle powiązana z teorią zbiorów rozmytych.

Ogromny wpływ na jej powstanie mieli polscy matematycy/logicy: Jan Łukasiewicz i Alred Tarski.

Logika rozmyta - przykład



Systemy rekomendacyjne

Silniki rekomendacji

allegro

Systemy rekomendacyjne - metody filtrowania

Content based filtering

Filtrowanie w oparciu o indywidualne preferencje użytkownika

Collaborative filtering

Filtrowanie w oparciu o preferencje użytkowników o podobnym guście.

Rozwiązania hybrydowe

Filtrowanie łączące obie metody.

Content based filtering - wady i zalety

Brak problemu "zimnego startu"

Szybkość

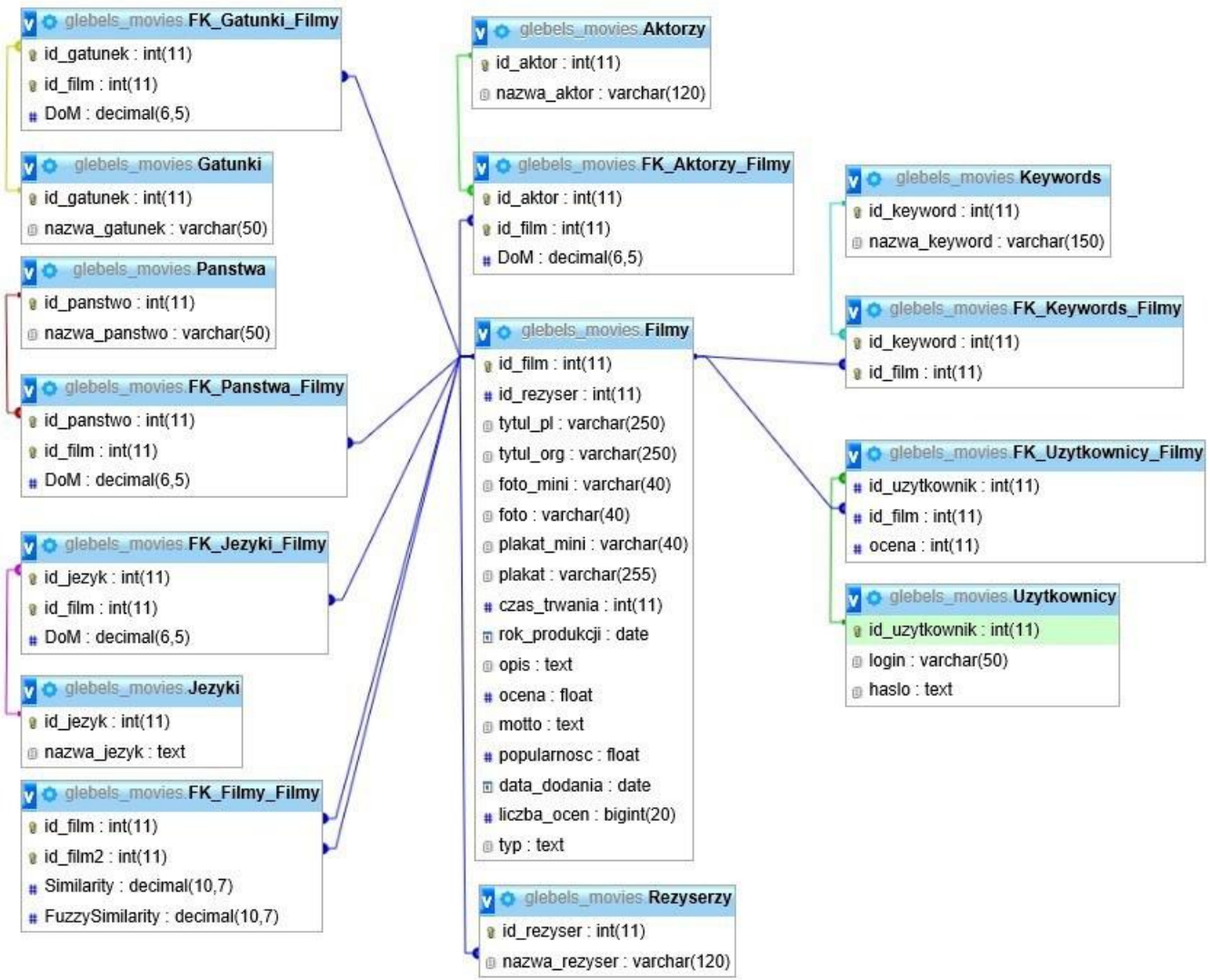
Daje gorsze rezultaty niż CF *

Przykład silnika rekomendacji

Silnik rekomendacji filmów z filtrowaniem opartym o logikę rozmytą

- 2727 filmów.
- 15986 aktorów.
- 6563 słów kluczowych.
- 1633 reżyserów.
- 108 języków.
- 92 państw.
- 24 gatunków.

Struktura bazy danych



Etapy budowania silnika rekomendacji

1. Filtrowanie filmów lubianych przez daną osobę.
2. Filtrowanie atrybutów filmów.
3. Wyznaczenie stopnia podobieństwa pomiędzy filmami.
4. Wyznaczenie współczynnika wsparcia rekomendacji

Etapy budowania silnika rekomendacji

1

Filtrowanie fimów lubianych przez daną osobę.

2

Filtrowanie atrybutów filmów.

3

Wyznaczenie stopnia podobieństwa pomiędzy filmami.

4

Wyznaczenie współczynnika wsparcia rekomendacji

Kiedy możemy uznać, że użytkownik polubił dany film?

$$\mu_L(x_j) > 0.5$$

$$\mu_L(x_j) = \frac{o - min}{max - min}$$

gdzie:
o - ocena użytkownika
min - minimalna ocena w danej skali
max - maksymalna ocena w danej skali

Kiedy możemy uznać, że użytkownik polubił dany film? - Przykład nr. 1.

W skali 1-5 użytkownik ocenił "film A" na 4.

$$\mu_L(x_A) = \frac{o - min}{max - min}$$

$$\mu_L(x_A) = \frac{4-1}{5-1} = \frac{3}{4} = 0.75$$

$$\mu_L(x_A) > 0.5$$

Stopień przynależności jest większy niż 0.5, a więc film można zaliczyć do zbioru filmów lubianych przez danego użytkownika.

Kiedy możemy uznać, że użytkownik polubił dany film? - Przykład nr. 2.

W skali 1-5 użytkownik ocenił "film B" na 3.

$$\mu_L(x_B) = \frac{o - min}{max - min}$$

$$\mu_L(x_B) = \frac{3-1}{5-1} = \frac{2}{4} = 0.5$$

$$\mu_L(x_B) = 0.5$$

Stopień przynależności jest równyniż 0.5, a więc filmu nie można zaliczyć do zbioru filmów lubianych przez danego użytkownika.

Filmy ocenione przez Dominika

Etapy budowania silnika rekomendacji

1

Filtrowanie fimów lubianych przez daną osobę.

2

Filtrowanie atrybutów filmów.

3

Wyznaczenie stopnia podobieństwa pomiędzy filmami.

4

Wyznaczenie współczynnika wsparcia rekomendacji

Filtrowanie atrybutów filmów

Obliczenie stopnia przynależności filmu do danego gatunku/aktora/języka/kraju produkcji.

$$\mu_{x_j}(A_{j_i}) = \frac{e-(i-1)}{\sum_{k=1}^{e} k}$$

gdzie:

x_j - j-ty film z listy filmów

i - numer atrybutu z listy A_j

e - liczba wszystkich atrybutów listy A_j

A_{j_i} i-ty element z j-tej listy atrybutów danego filmu

Filtrowanie atrybutów filmów - "The Goodfather"

Tytuł	Gatunek
The Godfather	Crime
The Godfather	Drama

$$\mu_{The\ Godfather}(A_{Crime}) = \frac{2-(1-1)}{\sum_{k=1}^{2} k} = \frac{2}{3} = 0.667$$

$$\mu_{The\ Godfather}(A_{Drama}) = \frac{2-(2-1)}{\sum_{k=1}^{2} k} = \frac{1}{3} = 0.333$$

Filtrowanie atrybutów filmów - "Goodfellas"

Tytuł	Język
Goodfellas	English
Goodfellas	Italian

$$\mu_{Goodfellas}(A_{English}) = \frac{2-(1-1)}{\sum_{k=1}^{2} k} = \frac{2}{3} = 0.667$$

$$\mu_{Goodfellas}(A_{Italian}) = \frac{2-(2-1)}{\sum_{k=1}^{2} k} = \frac{1}{3} = 0.333$$

Filtrowanie atrybutów filmów - "Casino"

Tytuł	Gatunek
Casino	Biography
Casino	Crime
Casino	Drama

$$\mu_{Casino}(A_{Biography}) = \frac{3-(1-1)}{\sum_{k=1}^{3} k} = \frac{3}{6} = 0.5$$

$$\mu_{Casino}(A_{Crime}) = \frac{3-(2-1)}{\sum_{k=1}^{3} k} = \frac{2}{6} = 0.333$$

$$\mu_{Casino}(A_{Drama}) = \frac{3-(3-1)}{\sum_{k=1}^{3} k} = \frac{1}{6} = 0.167$$

Etapy budowania silnika rekomendacji

1

Filtrowanie fimów lubianych przez daną osobę.

2

Filtrowanie atrybutów filmów.

3

Wyznaczenie stopnia podobieństwa pomiędzy filmami.

4

Wyznaczenie współczynnika wsparcia rekomendacji

Wyznaczanie stopnia podobieństwa pomiędzy filmami

Dwie podstawowe metody:

- *Fuzzy Set Theoretic*

$$S(x_j, x_k) = \frac{\sum_{i=1}^{n} min(\mu_{x_j(A_i)}, \mu_{x_k(A_i)})}{\sum_{i=1}^{n} max(\mu_{x_j(A_i)}, \mu_{x_k(A_i)})}$$

- *Fuzzy Theoretic Cosine*

$$S(x_j, x_k) = \frac{\sum_{i=1}^{n} \mu_{x_j(A_i)} \times \mu_{x_k(A_i)}}{\sqrt{(\sum_{i=1}^{n} (\mu_{x_j} A_j)^2)} \times \sqrt{(\sum_{i=1}^{n} (\mu_{x_j} A_k)^2)}}$$

Wyznaczanie stopnia podobieństwa pomiędzy filmami

Ostatecza forma wzoru na stopień podobieństwa pomiędzy filmami miała postać:

$$S(x_j, x_k) = S_{Gatunki}(x_j, x_k) \times 0.6 + S_{Aktorzy}(x_j, x_k) \times 0.2 + S_{Jezyki}(x_j, x_k) \times 0.1 + S_{Panstwa}(x_j, x_k) \times 0.1$$

Wyznaczanie stopnia podobieństwa pomiędzy filmami - Przykład: "The Godfather" vs "Casino"

Tytuł	Gatunek	DoM
The Godfather	Crime	0.667
The Godfather	Drama	0.333
Casino	Biography	0.5
Casino	Crime	0.333
Casino	Drama	0.167

Wyznaczanie stopnia podobieństwa pomiędzy filmami - Przykład: "The Godfather" vs "Casino"

Fuzzy Set Theoretic

$$S_{Gatunki}(A,B)=\frac{min(0,0.5)+min(0.667,0.333)+min(0.333,0.167)}{max(0,0.5)+max(0.667,0.333)+max(0.333,0.167)}=$$
$$=\frac{0+0.333+0.167}{0.5+0.667+0.333}=\frac{0.5}{1.5}=0.333$$

Fuzzy Theoretic Cosine

$$S_{Gatunki}(A,B)=\frac{(0\times0.5)+(0.667\times0.333)+(0.333\times0.167)}{(\sqrt{(0^2+0.667^2+0.333^2)})\times(\sqrt{(0.5^2+0.333^2+0.167^2)})}=$$
$$=\frac{0+0.22211+0{,}05561}{0{,}74551\times0{,}62352}=\frac{0{,}27772}{0{,}46484}=0{,}597$$

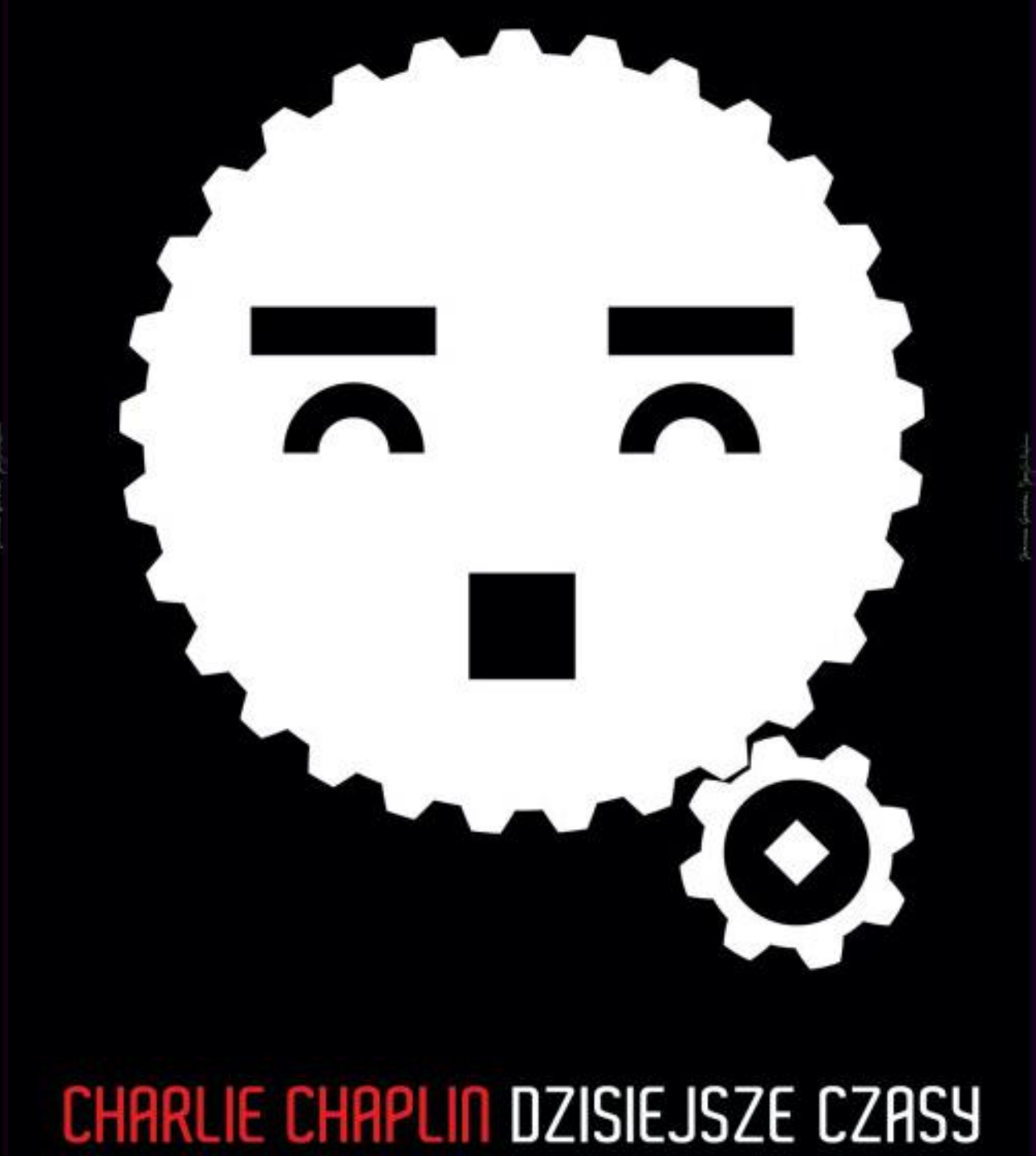
Wyznaczanie stopnia podobieństwa pomiędzy filmami

Tytuł	Tytuł	Stopień podobieństwa
Goodfellas	Casino	0.838
Goodfellas	The Godfather	0.587
The Godfather	Casino	0.512

Jakie filmy są do siebie najbardziej podobne?

	Casablanca	Chłopcy z ferajny	Czas Apokalipsy	Dwunastu gniewnych ludzi	Dzisiejsze czasy	Forrest Gump	Incepcja	Matrix	Milczenie owiec	Mroczny rycerz	Ojciec Chrzestny	Podziemny krąg	Poszukiwacze zaginionej arki	Psychoza	Pulp Fiction	Siedem	Skazani na Shawshank	Światła wielkiego miasta	Więzień nienawiści	Władca Pierścieni: Drużyna Pierścienia
Casablanca	1,00	0,35	0,69	0,65	0,39	0,75	0,16	0,16	0,43	0,33	0,38	0,64	0,17	0,17	0,43	0,17	0,39	0,52	0,39	0,11
Chłopcy z ferajny	0,35	1,00	0,36	0,41	0,29	0,39	0,18	0,18	0,60	0,45	0,59	0,40	0,16	0,22	0,58	0,48	0,58	0,31	0,58	0,12
Czas Apokalipsy	0,69	0,36	1,00	0,71	0,41	0,65	0,16	0,19	0,46	0,35	0,48	0,70	0,16	0,17	0,46	0,17	0,41	0,46	0,41	0,11
Dwunastu gniewnych ludzi	0,65	0,41	0,71	1,00	0,47	0,74	0,17	0,19	0,52	0,40	0,45	0,79	0,17	0,23	0,50	0,20	0,47	0,52	0,47	0,13
Dzisiejsze czasy	0,39	0,29	0,41	0,47	1,00	0,44	0,17	0,19	0,34	0,28	0,30	0,46	0,17	0,20	0,32	0,20	0,32	0,80	0,32	0,13
Forrest Gump	0,75	0,39	0,65	0,74	0,44	1,00	0,17	0,19	0,49	0,37	0,42	0,73	0,17	0,20	0,47	0,20	0,44	0,56	0,44	0,13
Incepcja	0,16	0,18	0,16	0,17	0,17	0,17	1,00	0,70	0,19	0,48	0,15	0,16	0,65	0,32	0,18	0,32	0,17	0,17	0,17	0,61
Matrix	0,16	0,18	0,19	0,19	0,19	0,19	0,70	1,00	0,19	0,52	0,17	0,18	0,73	0,19	0,17	0,19	0,19	0,19	0,19	0,69
Milczenie owiec	0,43	0,60	0,46	0,52	0,34	0,49	0,19	0,19	1,00	0,59	0,75	0,51	0,17	0,24	0,78	0,63	0,77	0,37	0,77	0,13
Mroczny rycerz	0,33	0,45	0,35	0,40	0,28	0,37	0,48	0,52	0,59	1,00	0,55	0,39	0,54	0,21	0,57	0,49	0,59	0,30	0,57	0,47
Ojciec Chrzestny	0,38	0,59	0,48	0,45	0,30	0,42	0,15	0,17	0,75	0,55	1,00	0,44	0,15	0,18	0,74	0,61	0,78	0,32	0,78	0,12
Podziemny krąg	0,64	0,40	0,70	0,79	0,46	0,73	0,16	0,18	0,51	0,39	0,44	1,00	0,16	0,19	0,49	0,24	0,46	0,51	0,51	0,13
Poszukiwacze zaginionej arki	0,17	0,16	0,16	0,17	0,17	0,17	0,65	0,73	0,17	0,54	0,15	0,16	1,00	0,17	0,17	0,17	0,17	0,17	0,17	0,68
Psychoza	0,17	0,22	0,17	0,23	0,20	0,20	0,32	0,19	0,24	0,21	0,18	0,19	0,17	1,00	0,22	0,41	0,20	0,20	0,20	0,13
Pulp Fiction	0,43	0,58	0,46	0,50	0,32	0,47	0,18	0,17	0,78	0,57	0,74	0,49	0,17	0,22	1,00	0,61	0,75	0,35	0,75	0,12
Siedem	0,17	0,48	0,17	0,20	0,20	0,20	0,32	0,19	0,63	0,49	0,61	0,24	0,17	0,41	0,61	1,00	0,67	0,20	0,63	0,13
Skazani na Shawshank	0,39	0,58	0,41	0,47	0,32	0,44	0,17	0,19	0,77	0,59	0,78	0,46	0,17	0,20	0,75	0,67	1,00	0,34	0,80	0,13
Światła wielkiego miasta	0,52	0,31	0,46	0,52	0,80	0,56	0,17	0,19	0,37	0,30	0,32	0,51	0,17	0,20	0,35	0,20	0,34	1,00	0,34	0,13
Więzień nienawiści	0,39	0,58	0,41	0,47	0,32	0,44	0,17	0,19	0,77	0,57	0,78	0,51	0,17	0,20	0,75	0,63	0,80	0,34	1,00	0,13
Władca Pierścieni: Drużyna Pierścier	0,11	0,12	0,11	0,13	0,13	0,13	0,61	0,69	0,13	0,47	0,12	0,13	0,68	0,13	0,12	0,13	0,13	0,13	0,13	1,00

Największe podobieństwo



Najmniejsze podobieństwo

Etapy budowania silnika rekomendacji

1. Filtrowanie fimów lubianych przez daną osobę.
2. Filtrowanie atrybutów filmów.
3. Wyznaczenie stopnia podobieństwa pomiędzy filmami.
4. **Wyznaczenie współczynnika wsparcia rekomendacji**

Wyznaczanie współczynnika wsparcia rekomendacji

$$\forall x_j \in N \; R(x_j) = \sum_{x_k \in L} \mu_L(x_k) \times S(x_j, x_k)$$

gdzie:

N - zbiór wszystkich filmów których użytkownik nie ocenił

$R(x_j)$ - funkcja wsparcia rekomendacji

L - zbiór filmów lubianych przez użytkownika

$mu_L(x_k)$ - stopień przynależności ocenionego filmu do zbioru filmów lubianych

$S(x_j, x_k)$ - funkcja zwracająca stopień podobieństwa pomiędzy filmami

Jakie filmy zarekomendujemy Dominikowi?

Podsumowanie

Pytania do Was

Która para filmów ocenionych przez Dominika była ze sobą najmocniej skoreowana?

Odp: "Chłopcy z ferajny" i "Kasyno"

Które para filmów z TOP 20 miała najmniejszy stopień podobieństwa?

Odp: Casablanba i Władca Pierścieni

Który film został zarekomendowany Dominikowi na pierwszym miejscu?

Odp: Donnie Brasco

Jakiego pojęcie używamy w teorii zbiorów rozmytych i logice rozmytej w alternatywie do prawdopodobieństwa.

Odp: Stopnia przynależności

Jaka jest główna zaleta Content Based Filtering?

Odp: Brak problemu "zimnego startu".

Pytania do mnie

Materiały: mateuszgrzyb.pl/WDS

Kontakt: m.grzyb@outlook.com